Point-Pair Descriptors for 3D Facial Landmark Localisation

Marcelo Romero and Nick Pears
Department of Computer Science
The University of York
York, UK
{mromero, nep}@cs.york.ac.uk

Abstract— **Our pose-invariant point-pair descriptors, which encode 3D shape between a pair of 3D points are described and evaluated. Two variants of descriptor are introduced, the first is the point-pair spin image, which is related to the classical spin image of Johnson and Hebert, and the second is derived from an implicit radial basis function (RBF) model of the facial surface. We call this a cylindrically sampled RBF (CSR) shape histogram. These descriptors can effectively encode edges in graph based representations of 3D shapes. Thus, they are useful in a wide range of 3D graph-based retrieval applications. Here we show how the descriptors are able to identify the nose-tip and the eye-corner of a human face simultaneously in six promising landmark localisation systems. We evaluate our approaches by computing root mean square errors of estimated landmark locations against our ground truth landmark localisations within the 3D Face Recognition Grand Challenge database.**

Keywords – 3D shape descriptors; 3D facial landmark localisation; 3D face alignment; invariance.

I. INTRODUCTION

Many face processing applications, such as face tracking, identification and verification require automatic landmark localisation. In the context of 3D face data, the nose is often quoted as the most distinctive feature [5], [6], [9] and, in addition, it is visible over a wide range of head poses. These facts make it an obvious landmark to include in a minimal set of three rigid features, which allow facial pose to be established and, if necessary, normalised [17]. One can also note that the inner eye corners are highly concave areas and so we use these additional two features to complete a minimal landmark triplet. A further reason that we have selected this landmark triplet is that it has been shown that this particular facial area is more distinctive for recognition using 3D data [13], and it has been proved robust in presence of facial expressions [10], [19].

In this paper, we introduce two variants of point-pair feature descriptors, which encode 3D shape between a pair of 3D points (candidate landmarks) in a pose invariant way.

The first is the point-pair spin image, which is related to the classical spin image of Johnson and Hebert [16], and the second is derived from an implicit radial basis function (RBF) model of the facial surface. We call this a cylindrically-sampled RBF (CSR) shape histogram. This is related to our previous work on spherically sampled RBF (SSR) shape histograms [17]. Both of these descriptors can effectively encode edges in graph based representations of 3D shapes, and are designed to be pose-invariant. Thus they are useful in a wide range of 3D graph-based retrieval applications, not just 3D face recognition.

Here, however, as a first application of these descriptors, we evaluate their ability to localise the nose tip and eye corner in a pose invariant way. This is possible by applying a two steps process: Firstly, we populate a pair of candidate landmark lists, using simple descriptors that measure local convexity. These descriptors are *distance to local plane* and the *SSR convexity values*, described in our previous work [20]. Then, candidate landmark pairs are selected as those landmarks that are within a trained Euclidean distance of one another. Finally, candidate point-pair features are generated and compared against trained point-pair descriptors in order to select the best landmark pair. Using this procedure, we have experimented six landmark localisation systems using our point-pair descriptors. In total, we have generated a manual mark up of eleven facial landmarks across the entire Grand Challenge 3D dataset [18]. Our future aim is to extend our landmark localisation processes to the full set of eleven features.

A. Related work

Automatic facial landmark localisation using only 3D data is attracting interest in the research community. Conde et al. [2] use spin images and support vector machine SVM classifier to locate the nose and the eyes. Xu et al. [5] present a 3D nose tip hierarchical filtering approach constructed with an *effective energy sphere* and SVM classifier. Colbry et al. [8] use shape index for anchor point localisation. Segundo et al. [6] report a successful localisation result experimenting with the FRGC database, although it is a technique constrained to a facial frontal pose. Lu and Jain [7] propose a feature extractor based on the directional maximum to estimate the nose tip location and pose angle simultaneously. Koudelka et al. [21] used radial symmetry and shape to extract features of interest on 3D range images of faces. Different pose-dependent and multimodal approaches to localise facial landmarks using the FRGC database have been also reported [3], [10]-[12].

Our final objective is a landmark localisation approach robust to extreme pose variations, which is relevant to unconstrained face recognition [1], [4] and the 3D face recognition challenge [14]. Results presented here are motivating and guiding our future work towards that final objective.

978-1-4244-XXXX-X/09/$25.00 ©2009 Crown

B. Preliminaries

In previous experimentation, we have reported facial landmark localisation using our structural graph matching algorithm [18], which is a 'relaxation by elimination' implementation, where simple descriptors are matched to localise the eye-corners and the nose tip simultaneously. A more sophisticated approach is our cascade filtering implementation [20], using either spin-images [16] or SSR histograms [17] to encode the local shape around a 3D surface point. Here the nose tip is robustly localised prior to localising the eye corners. For all of them, we have defined separate training and testing sets within the FRGC database. Firstly, we establish visually that 2D and 3D datasets are correctly registered. For those that are, we collect eleven ground-truth landmarks [18] by manually clicking 2D features on enlarged bitmaps, and mapping them to 3D (note that only three of these 3D landmarks are reported in this paper). We down-sample the data by a factor of 4, mapping 3D landmarks to the nearest down-sample. We then process this data in order to establish a variety of feature descriptors, to be used for training purposes.

After our feature localisation processes have finished, performance results are collected by comparing the localised feature landmarks against our ground-truth data. Since the definition of successful landmark localisation is dependent on setting a threshold of acceptable error, we explore performance over the full range of possible thresholds. This allows us to identify both gross errors ('fails'), where the system completely fails to identify the correct landmark, and errors of poor localisation, which are due to the combined effect of any inaccuracies in the system.

The rest of this paper is structured as follows. Previously reported vertex-based feature descriptors are briefly reviewed in section 2. Section 3 introduces our new point-pair spin images and CSR shape histograms. Our experimental framework is described in section 4, followed by the evaluation procedure and results in section 5. Finally, conclusions and future work are discussed in section 6.

II. VERTEX BASED FEATURE DESCRIPTORS

We have selected a range of feature descriptors with various computational costs of extraction and powers of discrimination in our facial landmark experimentations. These descriptors are pose-invariant, and robust to multi-resolution and poor quality raw data. We now provide a brief description of those used in our systems.

A. Distance to local plane (DLP)

To compute the DLP of a vertex, we determine the neighbour vertices that lie within some predetermined radius. We fit a least-squares plane to these points by applying singular value decomposition (SVD) to the mean-centered data points and we ensure that the plane normal always points towards the camera coordinate system. We can then apply a dot product operation to determine the distance of the candidate vertex to that local plane. By bounding the allowable values of DLP using the Mahalanobis metric, referenced to the mean and variance of our training data, we can identify both nose (suitably convex) and eye corner (suitably concave) candidate vertices.

B. Spherically sampled RBF (SSR) descriptors

An SSR shape histogram is a local surface representation derived from an implicit radial basis function (RBF) model, which is in fact a *signed distance to surface function* [17].

The concept of SSR shape histogram is to sample the RBF function in the locality of a candidate facial vertex in a pose-invariant way. Thus, a set of n sample points are evenly distributed across a unit sphere, centred on the origin. The sphere is then scaled by q radii, r_i, to give a set of concentric spheres and their common centre is translated such that it is coincident with a facial surface point.

The RBF function, s, is then evaluated at the $N = nq$ sample points on the concentric spheres and these values are normalised by dividing by the appropriate sphere radius, r_i, giving a set of values in the range -1 to 1. If we bin the normalised values $s_n = \frac{s}{r_i}$ over p bins, we can construct a $(p \times q)$ SSR histogram.

Clearly, the convexity of the local surface shape around some point is related to the distribution of values within the SSR histogram. In [17], Pears shows that a fast way to approximately compute surface convexity is to use a single sampling sphere (we choose the radius as 20mm, which is related with the volume of the human nose), and compute convexity as the average of the signs of the RBF evaluations over that sphere. Thus a convexity value, C, is given as:

$$C = \frac{1}{n}\sum_{i=1}^{n} sign(s_i)$$

Note that this is much cheaper to compute than SSR histograms, as there are no binning operations, but we have a lower dimensional, less discriminating feature.

III. POINT-PAIR FEATURE DESCRIPTORS

A. Point-pair spin images

A point-pair spin image is a modification of Johnson and Hebert's classical spin image [16], which cylindrically encodes 3D shape around some specified surface point, relative to the surface normal of that point. In the point pair spin image representation, we define a direction using a pair of 3D surface points, which are landmark candidates in our application. Points lying within a 3D solid cylinder of some radius, and which has its length and axis defined by the 3D point pair, are binned into a two-dimensional histogram. One dimension of bins encodes a range of different radii from the 3D point-pair axis, and the other dimension of bins encodes normalised distances along the axis (we refer to this as a height), where the normalisation is achieved by dividing by the length of the cylinder axis. Note that this

descriptor is pose invariant, but is *directed*, in the sense that we encode shape in a consistent direction, from one 3D landmark to another. We can envisage different approaches and applications where one might wish to use an *undirected* descriptor, in which case, the distance along the cylinder axis should be measured from the centre and should be unsigned.

B. Cylindrically Sampled RBF (CSR) histograms

CSR histograms are analogously derived from Pears' SSR descriptors [17], as point-pair spin images are derived from classical spin images.

For CSR shape histograms, a cylindrical 3D sampling pattern is produced by generating a set of n sample points around each of q concentric circles. This set of q concentric circles is then repeated at regular intervals along the axis defined by the 3D point pair to gives h sets of concentric circles (we refer to these different axial positions as variations in heights along the sampling cylinders). This cylindrical sampling pattern, placed between the nose tip and left eye corner is shown in figure 1. Thus, the RBF, s, is evaluated at $N = nqh$ sample points on a set of concentric cylinders, and these evaluations are normalised by dividing by the associated cylinder radius, r_i, giving a set of values that mostly lie in the range -1 to 1. In our experiments here we set $h = q = 8$, which means that we have eight cylinders, with eight sampling planes at different heights on that cylinder. If we bin the normalised RBF evaluations $s_n = \frac{s}{r_i}$ over p bins, we can construct a $(p \times q)$ CSR shape histogram. Note that, in constructing such a histogram, we can bin relative to the 8 radii or the 8 normalised height values along the cylinder, or we could retain all information in a $(p \times q \times h)$ histogram. We investigate these three approaches in our experimentation. Figure 2 shows CSR histograms when binning against radii (top) and binning against heights (bottom).

IV. Experimental Framework

Our point-pair feature descriptors were experimented in six different systems as described in table I. Our experimental framework (illustrated in figure 3) is as follows: To extract a point-pair descriptor, a set of candidate pairs needs to be created initially. To do this, we use vertex-based feature descriptors (DLP and SSR values) which encode shape in a spherical neighbourhood of a *single* vertex, and have proved to be robust in previous experimentation [17], [18], [20].

For a given set of 3D point clouds, we compute DLP values and only points within three standard deviations from trained DLP data of the nose-tip and eye-corner are retained. Every DLP candidate point is now compared against trained SSR convexity values, and only candidate points below SSR value thresholds are retained.

We have now two lists of candidate points which have been evaluated with local (spherical neighbourhood) feature descriptors, and we have observed clusters of similar values around the nose tip and eye corner regions. However, evaluating every possible combination is computationally expensive and we seek to further reduce the number of candidates. We do this by removing the candidate vertices that are not a local minimum in Mahalanobis value, within some predefined spherical neighbourhood.

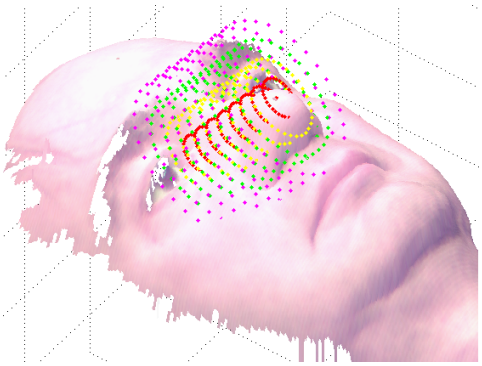

Figure 1. Cylindrical sampling pattern used to generate CSR histograms. This is shown positioned from the nose-tip to the left inner-eye corner, as occurs in both training and testing phases.

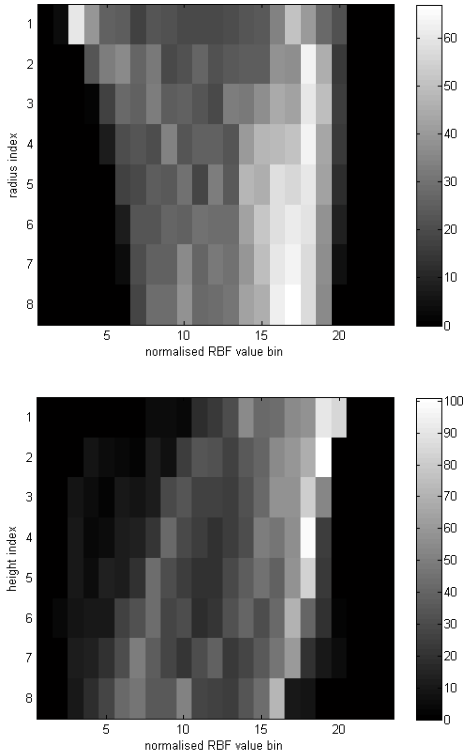

Figure 2. CSR histograms describing 3D shape from nose tip to the left eye corner. Top histogram has binned values with respect to radius and bottom histogram has binned values with respect to height along cylinder axis.

Then, pairs of candidates are produced by exhaustive combination, and we eliminate unlikely pairs by using trained Euclidean distance information between nose tips and

eye corners. Here, every pair of candidates within three standard deviations of trained Euclidean distance is retained. Next, for every pair of candidates a point-pair descriptor is computed and compared against trained point-pair data. Finally, the pair with minimum Mahalanobis distance to the trained point-pair mean vector is stored for performance evaluation.

TABLE I. IMPLEMENTATIONS USING POINT-PAIR DESCRIPTORS

	Point-pair descriptor
System 1	$(p \times q)$ CSR histograms binned against radii.
System 2	$(p \times h)$ CSR histograms binned against height.
System 3	Similar as system 2, but using only a single radius (i.e. single cylinder, radius = 20 mm).
System 4	$(p \times q \times h)$ CSR shape histograms.
System 5	Directed point-pair spin images
System 6	Undirected point-pair spin images

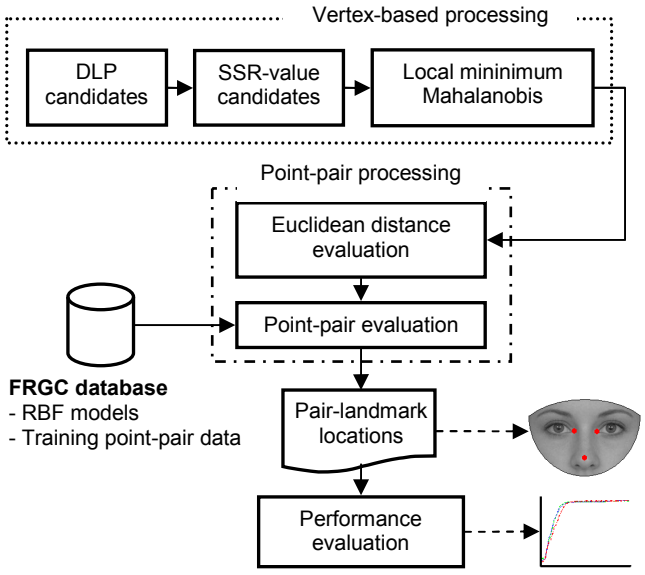

Figure 3. Block diagram of the complete point-pair experimentation for systems in table I.

V. EVALUATION

We have evaluated our point-pair descriptors on the FRGC 3D dataset [15]. Here we detail our methodology and results.

A. Methodology

Six landmark localisation systems have been created for the inner eye corners and the nose tip, each of which uses the same training and testing 3D scans. However, they use different point-pair descriptors as mentioned in table I. Our experimental methodology was as follows:

1. For each record in the FRGC database, eleven landmarks (we only use three here) were collected by very carefully manually clicking on enlarged intensity images and then computing the corresponding 3D point using the registered 3D shape information.
2. We selected the first 200 subjects from the Spring-2003 subset, which have more than one image in this specific data subset. For each of these persons, we randomly selected a capture to give 200 training 3D images.
3. For each of these 200 training 3D images, CSR shape histograms at the ground-truth nose tip to eye corner vertices were constructed, using 8 height values and 8 radii of 10 mm to 45 mm in steps of 5 mm and 23 bins for normalised RBF values. Systems 1 to 3 use CSR shape histograms of dimension (8×23), but they binned against radii and height as mentioned in table I. System 4 retain both radii and height in $(8 \times 23 \times 8)$ CSR shape histograms.
4. For the same training set above, point-pair spin images from the ground-truth nose tip to eye corner vertices were computed. In this method, spin-coordinates were calculated using the direction vector from the nose tip to the inner-eye vertices. Directed and undirected point-pair spin images were generated from the cylinder's centre. An (8×23) spin-image was produced using $\alpha_{\max} = 45mm$ and $\beta_{\max}$ equals the half of the total height of the cylinder, these values cover an equivalent volume as the CSR histograms do.
5. DLP and SSR values were computed, using a radius of 20 mm and 128 sample points for SSR values.
6. We evaluated our localisation systems in two scenarios, considering variations in depth and facial expressions (see table II). Naturally, there are variations in illumination and small variations in pose.
7. We applied principal component analysis (PCA) to reduce the point-pair descriptor feature space dimensionality to 64.
8. For all pair candidates on all test images, we calculated the Mahalanobis distance to the trained point-pair mean vector. For each test image, the pair of candidates with the minimum Mahalanobis distance was identified as the nose tip and eye corner, and then stored.
9. We gather results by computing the root mean square (RMS) error of the automatically localised landmarks with respect to the landmarks manually labelled in our ground truth. Remember that localisation is done at the 3D vertex level and we are using a down-sample factor of four on the FRGC dataset, which gives a typical distance between vertices of around 3-5 mm. This has implications on the achievable localisation accuracy. We set a distance threshold (specified in millimetres) and if the RMS error is below this threshold, then we label our result as a successful localisation. This allows us to present a performance curve indicating the percentage of successful feature localisations against

the RMS distance metric threshold used to indicate a successful location. These results have the nice property that they are not dependent on a single threshold and, in general, these performance curves show two distinct phases: (i) a rising phase where an increased RMS distance threshold masks small localisation errors, and (ii) a plateau in the success rate, where an increased RMS threshold does not give a significant increase in the success rate of localisation. If the plateau is not at 100% success rate, this indicates the presence of some gross errors in landmark localisation. A sensible place to choose the threshold is close to where the graph switches from the rising region to the plateau region. Of course, it is useful to choose some RMS threshold values and quote performance figures, e.g. categorisation in table III.

TABLE II. TESTING SETS FOR EVALUATION PERFORMANCE ANALYSIS

Scenarios	Subset	Size
1. Depth variations, neutral expressions.	Spring2003	509
2. Facial expression variations and few depth variations.	Fall2003	1,507
	Spring2004	1,764

TABLE III. THRESHOLDS TO EVALUATE ESTIMATED LOCATIONS

Success	$RMS \leq 12mm$
Poor	$12mm < RMS \leq 20mm$
Fail	$RMS > 20mm$

B. Results

Figure 4 shows the overall localisation performance for the inner-eye corners and the nose tip using our six landmark localisation systems (top row for the eye-corners and bottom row for the nose tip). These results were generated by averaging the results from the three data sets presented in table II. Table IV summarises localisation performance, where success is defined according to an RMS error threshold of 12 mm (see table III). These results indicate that the nose tip is a more distinctive landmark in 3D data in comparison with the eye corner. From table IV, one can observe that a histogram which bins against radii (system 1) has produced a slightly better result in comparison with our other three CSR histogram implementations. In this case, 99.65 % and 96.03 % of the nose tips and inner-eye corners respectively have been successfully localised. Undirected point-pair spin images (system 6) are reporting a better performance than the directed point-pair spin images (system 5), in the sense that 98.75% and 91.29% of the nose tips and inner-eye corners respectively were successfully localised.

VI. CONCLUSIONS

Our point-pair feature descriptors have been presented and experimented. These are novel pose-invariant feature descriptors which encode 3D shape between a pair of 3D points.

We have reported six landmark localisation systems using our point-pair descriptors to localise the inner-eye corner and the nose tip simultaneously. Our experimental results indicate their robustness to localise such distinctive facial landmarks in 3D data. System 1 (which bins against radii) has produced a better result than our other CSR histogram implementations (system 2–4). On the other hand, undirected point-pair spin images show a better performance than the directed version. Different factors are combined in these results, e.g. encoding method, number of heights, bins and radius, etc. We are including a further analysis of these factors as part of our future work.

Two testing scenarios have been experimented, considering depth and facial expression variations using the 3D FRGC dataset.

Our experimental framework uses both pose invariant feature descriptors and a spherical neighbourhood approach. This is a key difference against some other pose dependant localisation methods in the literature.

TABLE IV. SUCCESSFUL LOCALISATION SUMMARY

		Scenario #1	Scenario #2		Overall
		Spring-03	Fall-03	Spring-04	
System 1	Eyes	91.94 %	96.81 %	96.54 %	96.03 %
	Nose	99.60 %	99.53 %	99.77 %	99.65 %
System 2	Eyes	92.73 %	94.29 %	96.20 %	94.97 %
	Nose	99.60 %	99.20 %	99.60 %	99.44 %
System 3	Eyes	90.17 %	92.76 %	93.31 %	92.67 %
	Nose	99.01 %	98.73 %	99.37 %	99.07 %
System 4	Eyes	90.56 %	89.38 %	92.00 %	90.76 %
	Nose	98.23 %	97.54 %	98.07 %	97.88 %
System 5	Eyes	75.83 %	86.26 %	85.88 %	84.68 %
	Nose	95.28 %	98.00 %	98.52 %	97.88 %
System 6	Eyes	84.08 %	92.16 %	92.63 %	91.29 %
	Nose	96.26 %	98.87 %	99.37 %	98.75 %

ACKNOWLEDGMENT

M. R. is supported by the Mexican Council of Science and Technology (CONACYT, grant 207690) with assistance of the Mexican Department of Public Education (SEP) and Government.

REFERENCES

[1] S. K. Zhou, R. Chellappa and W. Zhao, Unconstrained Face Recognition. Springer, USA, 2007.

[2] C. Conde, L. J. Rodriguez, E. Cabello, "Automatic 3D Face Feature Points Extraction with Spin Images", Proc. International Conference on Image Analysis and Recognition, LNCS 4142, 317 – 328, 2006.

[3] C. Boehnen and T. Russ, "A Fast Multi-Modal Approach to Facial Feature Detection", Proc. IEEE Workshop on Applications of Computer Vision, 2005.

[4] W. Zhao and R. Chellappa, Face Processing: Advanced Modeling and Methods, Elsevier, USA, 2006.

[5] C. Xu, T. Tan, Y. Wang and L. Quan, "Combining Local Features for Robust Nose Location in 3D Facial Data", Pattern Recognition Letters 27, 2006, 1487-1494.

[6] M. Segundo, C. Queirolo, O. Bellon and L. Silva, "Automatic 3D Facial Segmentation and Landmark Detection", Proc. 14th Int. Conf. on Image Analysis and Processing, 2007, 431 – 436.

[7] X. Lu and A. Jain. “Automatic Feature Extraction for Multiview 3D Face Recognition”, Proc. 7th International Conference on Automatic Face and Gesture Recognition, 2006.

[8] D. Colbry, G. Stockman, and A. Jain, “Detection of Anchor Points for 3D Face Verification”, Proc. Computer Vision and Pattern Recognition, 2005.

[9] S. Gupta, K. Aggarwal, M. Markey and A. Bovik, “3D Face Recognition Founded on the Structural Diversity of Human Faces”, Proc. Computer Vision and Pattern Recognition, 2007.

[10] K. Chang, K. Bowyer and P. Flynn, “Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28 (10), Oct. 2006.

[11] A. Salah and L. Akarun, “3D Facial Feature Localization for Registration”, International Workshop on Multimedia Content Representation, Classification and Security, 2006.

[12] A. Mian, M. Bennamoun and R. Owens, “An Efficient Multimodal 2D-3D Hybrid Approach to Automatic Face Recognition”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29 (11), 2007, 1927 – 1943.

[13] T. Nagamine, T. Uemura and I. Masuda, “3D Facial Image Analysis for Human Identification”, Pattern Recognition 1992, vol. I. Conference A: Computer Vision and Applications, Proc. 11th IAPR International Conference on Volume I (9), 1992, 324 – 327.

[14] K. Bowyer, K. Chang and P. Flynn, “A Survey of Approaches and Challenges in 3D and Multi-modal 3D+2D Face Recognition”, Computer Vision and Image Understanding 101, 2006, 1-15.

[15] P. J. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek, “Overview of the Face Recognition Grand Challenge”, Proc. Computer Vision and Pattern Recognition, vol. 1, 2005, 947- 954.

[16] A. Johnson and M. Hebert, “Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes”, Trans. Pattern Analysis and Machine Intelligence, vol. 21, number 5, 1997, 433 – 449.

[17] N. Pears, “RBF Shape Histograms and Their Application to 3D Face Processing”, Proc. 8th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2008.

[18] M. Romero and N. Pears, “3D Facial Landmark Localisation by Matching Simple Descriptors”, Proc. IEEE Int. Conf. Biometrics: Theory, Applications and Systems, 2008.

[19] A. Bronstein, M. Bronstein and R. Kimmel, “Three-Dimensional Face Recognition”, International Journal of Computer Vision, vol. 64 (1), Aug. 2005.

[20] M. Romero and N. Pears, “Landmark Localisation in 3D Face Data”, Proc. IEEE Int. Conf. Advance Video and Signal Based Surveillance 2009, in press.

[21] M. L. Koudelka, M. W. Koch and T. D. Russ, “A Prescreener for 3D Face Recognition Using Radial Symmetry and Hausdorff Fraction”, Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2005.

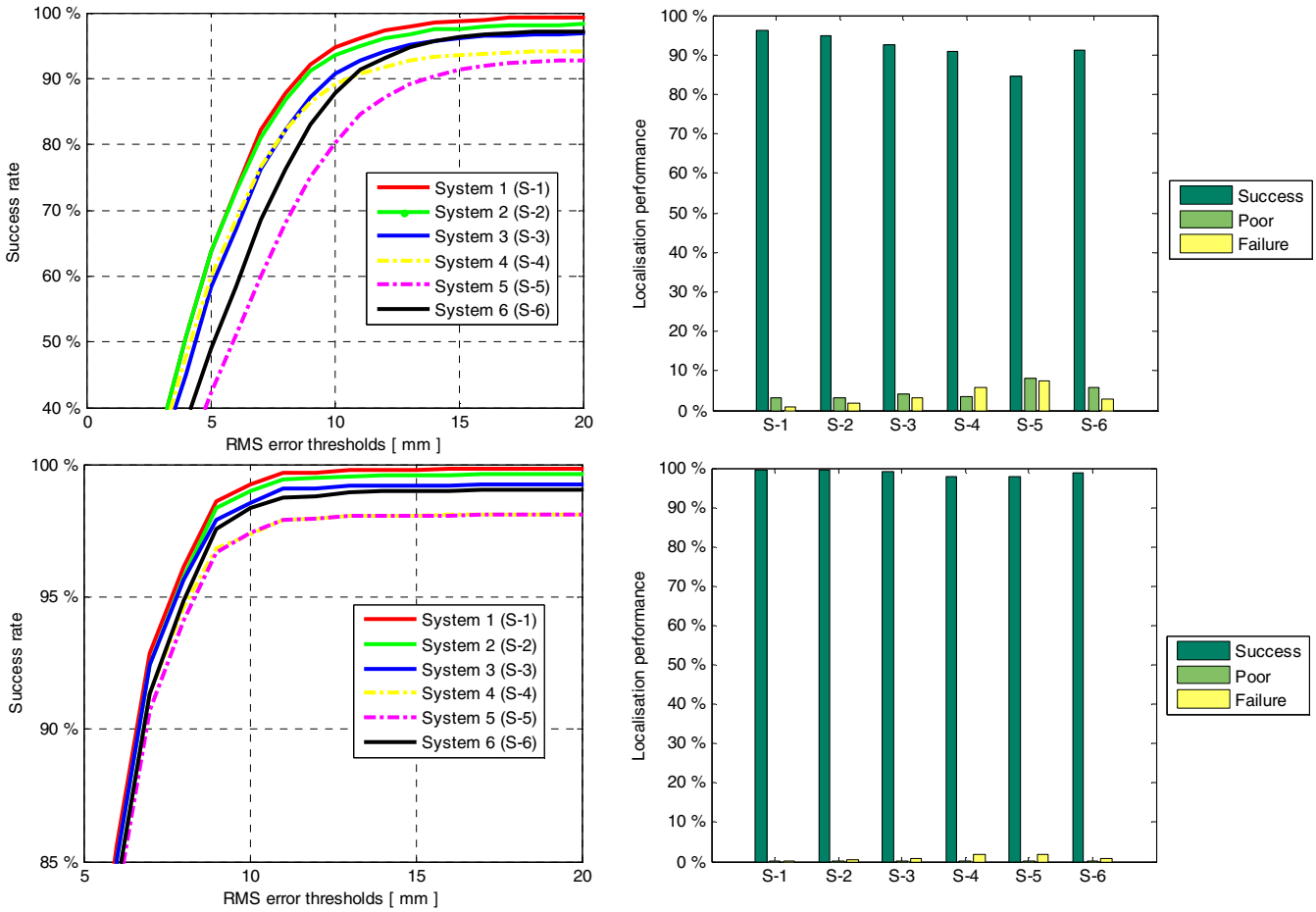

Figure 4. Localisation performance summary: for the inner-eye corner (top) and the nose tip (bottom). Cumulative RMS error curve (left) and overall localisation performance according to thresholds in table II (right).